G-42

教育用疑似 ID-POS データの作成方法の提案

Proposal for Creating Pseudo-ID-POS Data for Educational Purposes

磯部　貴翔 †　吉野　孝 †　大西　剛 ††　坂本　明一 ††　田井　紗瑛子 ††　南波　広哲 ††
Takato Isobe　Takashi Yoshino　Takeshi Onishi　Akikazu Sakamoto　Saeko Tai　Hiroaki Namba

1. はじめに

顧客ニーズの多様化により，顧客が何をどのように購入しているかを分析し，ニーズに合った商品やサービスを開発し，販売方法を考える必要があるため，近年は小売業のみならず取引先メーカーや卸売業も含め，ID-POS データの活用についての注目が高まっている．しかし，ID-POS データは今ではマーケティングにとって必要不可欠であるにもかかわらず，十分に活用できている企業は少ない．総務省の「デジタル・トランスフォーメーションによる経済へのインパクトに関する調査研究」[*1] では，34.5%の企業がデータを活用していないと回答しており，46.6%の企業がデータ活用で期待した効果が得られていないと回答している．このようなデータ活用における課題の原因として，データサイエンティストの人材不足が考えられる．特に日本では，データサイエンティストという職種が比較的新しいため実務経験のある人が少ないうえに，データ活用するために必要な知識やスキルを学べる機会が比較的少ない[*2]．さらに，企業が保持しているデータには顧客の個人情報が含まれていることが多く，コストや倫理上の観点からデータの入手が困難である．これらの課題の解決方法の一つとして，合成データ (現実世界のデータを模倣しているが実際のデータではない疑似的なデータ) の作成方法が用いられる．しかし，合成データ作成の多くは医療データを対象としたものであり，ID-POS データを対象とした研究は少ない．本研究では，合成データの考えに基づき，教育用疑似 ID-POS データの作成方法について検討を行う．ID-POS データのオープンデータ化は，多くの教育機関に提供を可能にするため，データ利活用コンペティションや統計教育などを行うことで，データサイエンス人材の育成支援につながると考えられる．

2. 関連研究

Chen らは，合成臨床データ生成ツールである Synthea の有用性を検証した [1]．Synthea によって生成された合成臨床データと現実世界のデータに関して，臨床品質指標を用いて比較したところ，Synthea の強みと弱みを明らかにし，特に臨床ガイドラインから逸脱した異質な介入後の健康アウトカムを現実的にモデル化するには限界があると報告した．この研究では ID-POS データを扱っておらず，臨床データに焦点を当てているため，対象とするデータの種類が異なる．また，本研究では疑似データの作成も行う．

高部は，公的統計ミクロデータの利用促進を目的として，法令および制度上の制約を満たしながら元データの構造を保持した教育用疑似ミクロデータの作成方法について検討を行った [2]．具体的には，企業に関する商用データを用いて，欧州で作成・提供されている合成データに関するモデルベースの手法を基に疑似ミクロデータを作成し，元データとの比較を行った．その結果，離散変数の分布，相関係数，売上高を予測する重回帰モデルの推定結果において，ほとんど同様の傾向を示した．この研究では，回帰モデルを用いたパラメトリックな手法であるが，本研究ではより高い精度を得るため，深層学習モデル (CTGAN) によるテーブルデータの生成方法を用いて疑似データの生成を提案する．

3. 疑似データの生成

3.1 方針

本研究では，疑似データの生成に CTGAN(Conditional Generative Adversarial Network)[3] を用いる．この手法は，GAN(Generative Adversarial Network) をベースとしたテーブルデータの生成モデルであり，生成されたデータは実際のデータではないため元データのプライバシー保護が可能である．そして，各変数を持った元データを CTGAN に学習させ，CTGAN により疑似データを生成する．

3.2 使用データ

POS(Point Of Sales) データとは，レジで商品のバーコードを読み取るなどをして取得したレシートデータである．ID-POS データとは，この POS データに顧客を識別する情報を紐づけたデータである．本研究で使用する ID-POS データは，株式会社オークワで記録されたデータであり，日付，時間，レシート番号[*3]，顧客コード[*4]，年代，性別，各商品の大分類，中分類，小分類，細分類，JAN コード[*5]，売上数量，単価の情報が記載されている．

商品は大分類，中分類，小分類，細分類に分類され[*6]，階層化されている．大分類では，農産，畜産，水産，惣菜，一般食品，菓子，酒，日配，米の 9 種類に分類されており，たとえば，大分類が畜産の場合，中分類には豚肉，鶏肉，国内牛肉，輸入牛肉，和牛，ミンチ，加工肉，畜産惣菜，関連食材の 9 種類に分類されている．中分類が和牛の場合の小分類には，スライス，しゃぶしゃぶ，切り落とし，ステーキ，焼肉，生肉，角煮・煮込，唐揚げ・カツ，ミンチ・ハンバーグ，その他を合わせた 10 種類に分類される．さらに小分類がスライスの場合の細分類は，和牛セット，和牛・ロイン・ヒレ，和牛・モモ，和牛・肩ロース，和牛・バラ，和

† 和歌山大学，Wakayama University
†† 株式会社オークワ，Okuwa Co., Ltd.

[*1] デジタル・トランスフォーメーションによる経済へのインパクトに関する調査研究の請負 (2021 年 3 月)，https://www.soumu.go.jp/johotsusintokei/linkdata/r03_02_houkoku.pdf

[*2] データサイエンティストが不足している理由とは?将来の需要も解説 (2023 年 9 月 18 日)，https://datamix.co.jp/media/careerenhancement/data-scientist/reason-for-shortage/

[*3] レシートに 1 枚ごとに付けられた番号

[*4] レシート番号に紐づき，会員である顧客を識別するための番号であり，匿名加工されている

[*5] 商品を識別するために付けられた番号

[*6] 大分類・中分類・小分類・細分類の順に細かくなっている

牛・ホルモン，和牛・その他，和牛・経産の 8 種類に分類される．

3.3 前処理

今回用いたデータでは,「JAN コード」「商品名」「単価」が含まれており，これらは商品を特定する可能性がある．商品の単価は店舗によって異なるため，店舗の売上実績に関する情報保護の観点から，これらの変数に以下の処理を行う．

- JAN コード (離散変数)
  「JAN コード」は生成した疑似データの精度に影響を及ぼすとは考えにくく，代表的な ID-POS データ分析に不必要な変数であるため削除する．

- 商品名 (離散変数)
  「商品名」はデータ分析に必要な変数であるため，細分類の名称にランダムな英数字で構成された文字列 3 桁[*7]を加えたものに変更する．

- 単価 (連続変数)
  「単価」はデータ分析に必要な変数であるため，5 円刻みとし，その商品の中央値の下一桁を 0 か 5 の近い方の値に変更する．しかし,「商品名」と「単価」を同時に学習させると，合成データ生成モデルの性質上，元データの「単価」の何十倍にも高くなったり低くなったりすることがあるため，学習の際は「単価」を削除し，疑似データ生成後に元データを参照しながら「商品名」に対して「単価」を結合する．

他の変数についても，以下のように前処理を行う．

- 日付 (連続変数)
  「日付」はデータ分析に必要な変数であるが，本研究では一日分のデータを使用することにより[*8],「日付」を学習する必要がなくなるため，学習の際は「日付」を削除し，疑似データの生成後に結合する．

- 時間 (連続変数)
  「時間」はデータ分析に必要な変数であるが,「時間」と「顧客コード」を同時に学習させると，合成データ生成モデルの性質上，一つの「顧客コード」に対し複数の「時間」が生成される可能性がある．これは，一日に複数回来店することを意味するが，現実的ではない間隔での来店かつ購買量が生成されるといった問題が考えられるため,「時間」は削除する．そして，疑似データの生成後に元データを参照しながら「顧客コード」に対して「時間」を結合する．

- レシート番号 (離散変数)
  「レシート番号」はデータ分析に必要な変数であるが,「レシート番号」と「顧客コード」を同時に学習させると，合成データ生成モデルの性質上，一つの「レシート番号」に複数の「顧客コード」が存在するといった問題が考えられる．仮に「レシート番号」を削除しても一日分のデータを使用する場合,「顧客コード」が「レシート番号」の代わりになるため,「レシート番号」は削除する．ここで，一日に複数回来店する顧客が考えられるが，そのような顧客は稀で[*9]，考慮する必要がないと考えられる．

- 顧客コード (離散変数)
  「顧客コード」はデータ分析に必要な変数であり，ここでは匿名化されているため，秘匿性が保たれていることから，そのまま学習に使用しても問題ないと考えられる．

- 年代 (離散変数)
  「年代」はデータ分析に必要な変数であり，ここでは 20 代，30 代，40 代，50 代，60 代，70 代，80 代の 7 つに区分しているため，秘匿性が保たれていることから，そのまま学習に使用しても問題ないと考えられる．しかし,「年代」と「顧客コード」を同時に学習させると，合成データ生成モデルの性質上，同一の「顧客コード」に対し異なる「年代」が生成されると考えられる．そこで,「年代」は削除し，疑似データの生成後に元データを参照しながら「顧客コード」に対して「年代」を結合する．

- 性別 (離散変数)
  「性別」はデータ分析に必要な変数であるが,「年代」と同様に,「顧客コード」と学習させると，合成データ生成モデルの性質上，同一の「顧客コード」に対し異なる「性別」が生成されると考えられる．そこで,「性別」は削除し，疑似データの生成後に元データを参照しながら「顧客コード」に対して「性別」を結合する．

- 分類 (離散変数)
  各商品の「大分類」「中分類」「小分類」「細分類」は生成した疑似データの精度に影響を及ぼすとは考えにくく，代表的な ID-POS データ分析に不必要な変数であるため削除する．しかし,「商品名」に使われている「細分類」の名称では「その他」などが複数存在するため，補足として，疑似データの生成後に元データを参照しながら「商品名」に対して「中分類」の名称を結合し，その変数名を「カテゴリ」に変更する．

- 売上数量 (連続変数)
  「売上数量」はデータ分析に必要な変数であり，そのまま学習に使用しても問題ないと考えられる．

学習に用いる変数は,「顧客コード」「商品名」「売上数量」の 3 つとする．

なお，今回使用したデータでは，欠損値は存在しないものの「年代」と「性別」で「不明」が存在する．同様に,「顧客コード」では「非会員」も存在する．これらは，RFM 分析やデシル分析などの顧客分析を行う際に不要であるため今回は削除している．

[*7] 文字列の重複はないことから，本研究で使用する ID-POS データの商品名が 238,328 種類以下であるため，商品名の重複をなくすことが可能

[*8] 他の日付の一日分の疑似データと結合することで，一週間，一か月，一年分の疑似データが作成可能だと考えられる

[*9] 主婦・主夫 500 人に聞いた「毎日スーパーで買い物する」人の割合は?(2023 年 8 月 16 日), https://news.mynavi.jp/article/20230816-2745332/

3.4 生成結果

ここで，生成するレコード数は元データと比較するため同じにする．生成前の元データの例を表 1，生成後の疑似データの例を表 2 にそれぞれ示す．元データを見ると，学習時に「顧客コード」がまとめられているにも関わらず，生成後の疑似データでは「顧客コード」が分散した状態であることが分かる．同様に,「商品名」も分散した状態で生成されている．

表 1: 生成前の元データの例

	顧客コード	商品名	売上数量 (個)
0	J9UW20FFGKS	食パン HIP	1
1	POMP90GFECS	もち米 Yg7	1
2	POMP90GFECS	焼菓子・駄菓子 H84	1
3	POMP90GFECS	少量菓子 dln	1
4	POMP90GFECS	スピリッツ Vzf	1
...	...	...	...
15781	P9RM10GKKHS	ミニトマト rHt	1
15782	P9RM10GKKHS	もぎたて生産者１ kHc	1
15783	P9RM10GKKHS	もぎたて生産者１ kHc	1
15784	P9RM10GKKHS	レギュラーコーヒー YZr	1
15785	P9RM10GKKHS	豆腐 4sl	1

表 2: 生成後の疑似データの例

	顧客コード	商品名	売上数量 (個)
0	K8TT50ECFFS	煮魚 OzP	1
1	P4NP10ADKDS	バナナ qhl	1
2	P7ML90FHEGS	シュレッド Yj0	2
3	P8UU70FABFS	板コンニャク 2wX	1
4	P6QP00FGKGS	カップめん NIK	1
...	...	...	...
15781	P6SR00GJKAS	麦 Yhs	1
15782	P8SL30FHDHS	薩摩もち豚 C39	1
15783	P5QL00GKDGS	きゅうり jj0	4
15784	P2RT90AACAS	赤 6wr	1
15785	P7SW30AJDFS	一本物 (ハーフ)WX7	5

4. 疑似データの作成結果と考察

作成した疑似データの有用性を確認するために，元データとの度数分布などの比較を行う．

4.1 疑似データの作成

疑似データの生成後，3.3 節で述べた各変数を結合する．表 3 に疑似データ作成後の例を示す．この表は，前節で述べた「顧客コード」が分散していた問題について,「顧客コード」を基準に並び替えを行った．その結果，ID-POS データとして問題ないように見える．

4.2 度数分布の比較

元データと疑似データにおいて,「顧客コード」「性別」「年代」「商品名」について集計した結果を表 4 に示す[*10]．この表から，特に「商品名」で大きな違いが見られたが，その他の変数ではほとんど等しいことが分かる．また,「カテゴリ」についての度数分布をグラフ化したものを図 1，図 2 にそれぞれ示す[*11]．これらの図を見ると，特に上位 3 個の野菜，パン，飲料の順序に違いがあるが，上位 11 個のカテゴリの要素の種類は同じであることが分かる．

4.3 統計分析結果の比較

ID-POS データにおける主なデータ分析のうち，ABC 分析とデシル分析を行った．ABC 分析とは商品を売上金額順に並べて A,B,C にグループ分けする分析であり，デシル分析とは顧客を購入金額順に並べて 10 分割にグループ分けする分析である．また，表 5 と表 6 に ABC 分析, 表 7 と表 8 にデシル分析の結果をそれぞれ示す[*12]．ABC 分析の結果では売上金額の高い上位 5 つを比較したところ，大きく差異が見られた．しかし，下位 5 つを比較すると，元データの結果と同様の傾向にあることが分かる．一方で，デシル分析の結果においては，上位と下位それぞれ 5 つの中で共通している顧客コードはなく，大きな差異が見られた．

4.4 考察と今後の課題

デシル分析におけるクラスの累積構成比について比較したものを表 9 に示しており，クラス「10」の累積構成比率の違いが大きいことが分かる．そこで，疑似データで購入金額の高かった顧客の購買履歴を見たところ，元データの最高購入金額者よりも 10 倍以上の購入金額を持つ人が疑似データに一定数存在することが分かった．

表 10 は疑似データの最高購入金額者の購買履歴の一部を示しているが，この表を見ると,「売上数量」が 8 であるにもかかわらず，ある商品のレコードが大量に生成されていることが分かる．このことが ABC 分析やデシル分析の結果に大きく影響を与えていたと考えられる．また，この事例以外にも同じことが起きていると考えられる.「売上数量」を考慮した上で，同時購入をモデル化することが今後，必要であると考えられる．一方で，4.2 節の結果を踏まえると，度数分布には問題ないと考えられる．

今回の結果より，疑似データの精度を向上させるために,「売上数量」は同時購入（レコード間の関係）について学習可能かということを検証・検討する必要があると考えられる．また，疑似データの有用性についての適切な評価を考えることも今後の課題である．

5. おわりに

本研究では，CTGAN を用いて疑似 ID-POS データを作成した．今回，ABC 分析などの統計分析の結果において元データと比較した際に差異が見られたため，今後は学習モデルと疑似データの評価を再検討し，精度の向上を目指す．同時に，疑似データの安全性についても評価していきたい．

参考文献

[1] Chen Junqiao, Chun David, Patel Milesh, Chiang Epson, James Jesse: The validity of synthetic clinical data: a validation study of a leading synthetic data generator

[*10] 秘匿性の観点から，詳細な数値は示していない

[*11] 秘匿性の観点から，縦軸の目盛りの数値は示しておらず，上位 15 個のみ示す

[*12] 秘匿性の観点から，売上 (購入) 金額は示していない

表 3: 疑似データ作成の例

	日付	時間 (時台)	顧客コード	性別	年代	カテゴリ	商品名	単価 (円)	売上数量 (個)
0	20221121	21	S9WP90ACFBS	女性	50 代	パン	地元パン 0O0	220	1
1	20221121	21	S9WP90ACFBS	女性	50 代	野菜	カット野菜 WLe	150	1
2	20221121	21	S9WP90ACFBS	女性	50 代	バンダイ，ちびっこ	ちびっこ og0	15	1
3	20221121	21	S9WP90ACFBS	女性	50 代	果物	パイン XKb	375	1
4	20221121	21	S9WP90ACFBS	女性	50 代	飲料	炭酸 (瓶，PET)Eus	125	1
...	...	...	...	...	...	...	...	...	...
15781	20221121	20	J0LP00AHFBS	女性	70 代	パン	食卓＋菓子パン袋 hsZ	150	1
15782	20221121	20	J0LP00AHFBS	女性	70 代	デザート	ヨーグルト Jj1	110	1
15783	20221121	20	J0LP00AHFBS	女性	70 代	飲料	お茶 (PET)1gX	90	1
15784	20221121	20	J0LP00AHFBS	女性	70 代	寿司米飯	丼 6gx	390	1
15785	20221121	20	J0LP00AHFBS	女性	70 代	アイスクリーム	その他 sZ9	335	1

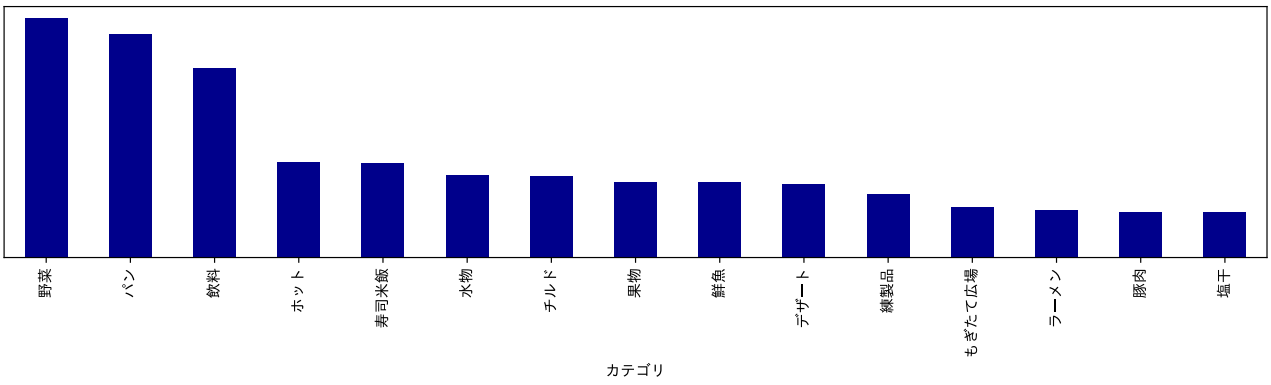

図 1: カテゴリの度数分布 (元データ)

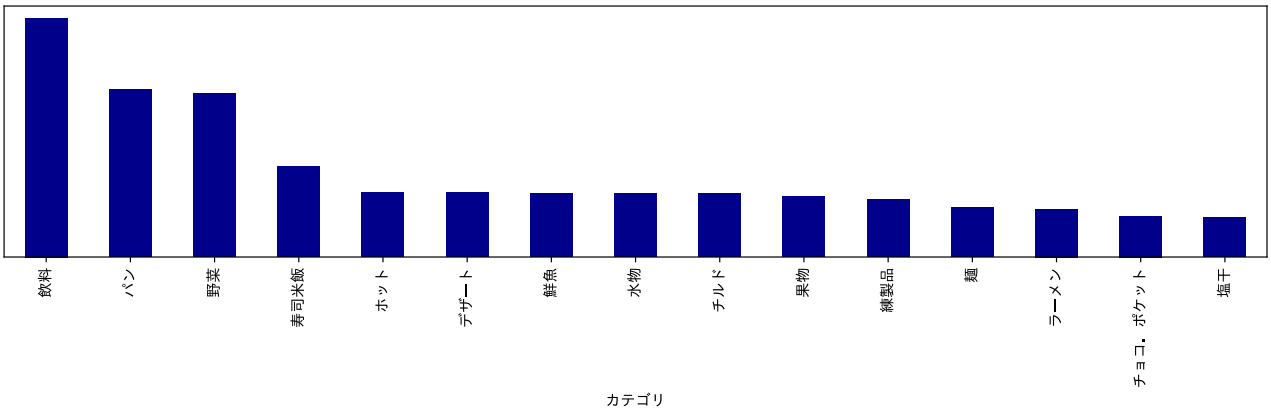

図 2: カテゴリの度数分布 (疑似データ)

表 4: 各変数の度数分布の比較

		相対誤差 (%)
顧客コード		0.39
性別	女性	0.29
	男性	0.71
年代	20 代	0.00
	30 代	0.00
	40 代	0.00
	50 代	0.73
	60 代	0.98
	70 代	0.18
	80 代	0.00
商品名		28.82

表 5: ABC 分析の結果 (元データ)

商品名	累積構成比 (%)	クラス
ミカン R3m	0.8	A
国内豚 VYO	1.4	A
牛乳 eKy	2.0	A
国内豚 ipp	2.5	A
こしひかり cd6	2.9	A
...	...	...
ちびっこ RPM	99.9	C
ちびっこ U3E	99.9	C
ちびっこ 23y	99.9	C
ちびっこ Dz5	99.9	C
パックガム 4r2	100.0	C

表 6: ABC 分析の結果 (疑似データ)

商品名	累積構成比 (%)	クラス
その他 ZwG	9.2	A
低脂肪・加工乳 (白)dbG	12.3	A
豆乳飲料 bcw	15.0	A
のむヨーグルト rAt	17.5	A
生ちくわ hkh	20.0	A
...	...	...
カット野菜サラダ IVM	99.9	C
ちびっこ 2rm	99.9	C
ちびっこ DFb	99.9	C
ちびっこ U3E	99.9	C
ちびっこ 2TN	100.0	C

表 7: デシル分析の結果 (元データ)

顧客コード	累積構成比 (%)	クラス
P8SL20FAEDS	0.6	10
P3WS90FHDES	1.1	10
P2LS30FGGES	1.5	10
P0WQ90FHDJS	1.9	10
P7MS40FFEDS	2.2	10
...	...	...
P6LS60FHEAS	99.9	1
P0PM20GEJFS	99.9	1
P5LQ00GHEGS	99.9	1
P8UL10GJKES	99.9	1
P6PU70BDDHS	100.0	1

表 8: デシル分析の結果 (疑似データ)

顧客コード	累積構成比 (%)	クラス
P6RN60GKCBS	7.1	10
P3NR30GKDDS	9.9	10
P2QW10FHEGS	12.5	10
P4QU90ADKES	14.9	10
P0LM70GDHCS	16.5	10
...	...	...
P1LL70FHKBS	99.9	1
P7NM00FHKGS	99.9	1
P4TS00GKEFS	99.9	1
P1UP90BFEJS	99.9	1
P4SQ60HFAJS	100.0	1

表 9: デシル分析におけるクラスの累積構成比

クラス	累積構成比 (%)	
	元データ	疑似データ
10	31.1	50.7
9	48.3	62.9
8	61.8	71.9
7	72.4	79.1
6	80.9	84.9
5	87.5	89.6
4	92.7	93.5
3	96.6	96.6
2	99.0	98.8
1	100.0	100.0

表 10: 疑似データの最高購入金額者の購買履歴の例

	日付	時間 (時台)	加工コード	性別	年代	カテゴリ	商品名	単価 (円)	売上数量 (個)
0	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
1	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
2	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
3	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
4	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
...	...	...	...	...	...	...	...	...	...
175	20221121	16	P6RN60GKCBS	女性	50 代	野菜	カット野菜 XbS	95	8
176	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
177	20221121	16	P6RN60GKCBS	女性	50 代	水物	国産納豆 B0Y	120	1
178	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8
179	20221121	16	P6RN60GKCBS	女性	50 代	花	その他 ZwG	345	8

(Synthea) using clinical quality measures，BMC medical informatics and decision making，19，pp.1–9 (2019).

[2] 高部 勲：合成データの考え方に基づく公的統計疑似ミクロデータの作成方法の検討，統計研究彙報= Research memoir of the statistics/総務省統計研修所 編，79，pp.111–129 (2022).

[3] Xu Lei，Skoularidou Maria，Cuesta-Infante Alfredo，Veeramachaneni Kalyan：Modeling Tabular Data using Conditional GAN，Advances in neural information processing systems，32，pp.1–11 (2019).